Міністерство освіти і науки України Національний технічний університет України "Київський політехнічний інститут імені Ігоря Сікорського"
Фізико-технічний інститут

КРИПТОГРАФІЯ
Комп’ютерний практикум №1
«Експериментальна оцінка ентропії на символ джерела відкритого тексту»

Виконали:

Факультет ФТІ,
група ФБ-93
Денисюк А.Г.
Товстоноженко М.С.

Київ-2021

Мета: Засвоєння понять ентропії на символ джерела та його надлишковості, вивчення та порівняння різних моделей джерела відкритого тексту для наближеного визначення ентропії, набуття практичних навичок щодо оцінки ентропії на символ джерела.

Завдання:

- Уважно прочитати методичні вказівки до виконання комп'ютерного практикуму.
- Написати програми для підрахунку частот букв і частот біграм в тексті, а також підрахунку H_1 та H_2 за безпосереднім означенням. Підрахувати частоти букв та біграм, а також значення H_1 та H_2 на довільно обраному тексті російською мовою достатньої довжини (щонайменше 1Мб), де імовірності замінити відповідними частотами. Також одержати значення H_1 та H_2 на тому ж тексті, в якому вилучено всі пробіли.
- За допомогою програми CoolPinkProgram оцінити значення $H^{(10)}$, $H^{(20)}$, $H^{(30)}$.
- Використовуючи отримані значення ентропії, оцінити надлишковість російської мови в різних моделях джерела

Хід роботи:

1)Створюємо допоміжні функції для таких цілей як: фільтрація вхідного тексту з файлу та приведення його до відповідного формату, розбиття на біграми (з заданим кроком), створення статистики по входженню біграм та символів в заданний текст, розрахунку частотності та запису результатів в файл. Завдяки цьому, спрощується процес розрахунку ентропії та роботи з текстом цілком.

Символи без пробілів:

symbol	frequency
'о'	0.10758271163013237
'а'	0.08257872445995995
'е'	0.07894153922489412
'и'	0.0731941072578311
'н'	0.0666315204999616
'т'	0.06136993260205679
'с'	0.05617922772943949
'л'	0.05121298577023267
'р'	0.04563980790700147
'в'	0.04398144046121222
'к'	0.033456787935709095
'м'	0.032195660777468646
'д'	0.031705386519229974
'п'	0.030761756245089873
'у'	0.02908714477267223
'ы'	0.02042907857974045
'г'	0.01945739044142404
'я'	0.0177089424843319
'б'	0.017001588961150196
'ь'	0.01662059270022978
'з'	0.016474396460574272
'ч'	0.012993744573018376

'й'	0.010955857596002197
'ж'	0.009614987034313292
'х'	0.009140956802703005
'ш'	0.006897804398291719 5
'ю'	0.005466853325299924
'ц'	0.003805532420123691
'э'	0.003733172665142681 7
'щ'	0.003671150018016102 3
'ф'	0.001509217746746764 4

Символи із пробілами:

symbol	frequency
' '	0.15315921440139063
'о'	0.09110542803368996
'а'	0.06993103189540359
'е'	0.06685091509357277
'и'	0.061983755291410564
'н'	0.05642628916581733
'т'	0.051970561936859484
'с'	0.04757486134472172
'л'	0.04336924510251424
'р'	0.03864965078253475
'в'	0.03724527759193142
'к'	0.028332572579081967
'м'	0.027264598665657884
'д'	0.02684941442765227
'п'	0.026050309824984838
'у'	0.024632180530110237
'ы'	0.017300176953523125
'г'	0.016477311807114408
'я'	0.014996654765552214
'б'	0.014397638952285076
'ь'	0.0140749957793771
'з'	0.013951190840935665
'ч'	0.011003632862082549
'й'	0.009277867053504993
'ж'	0.008142363173658311
'х'	0.007740935039923966
'ш'	0.005841342095554903
'ю'	0.004629554364749358
'ц'	0.003222680064278523 4
'э'	0.003161402872524683
'щ'	0.003108879565307105
'ф'	0.001278067142294393 2

Топ 15 біграм без пробілів з кроком 1:

symbol	frequency
'то'	0.014414087078278286
'ст'	0.013922244574223908
'на'	0.013093645040366534
'но'	0.011796835074721658
'ен'	0.011110027974465546
'ов'	0.01068021966011172
'по'	0.010440944928409591
'ко'	0.01013225098442351
'ал'	0.010042153708906041
'не'	0.009721643728786523
'ни'	0.009537018164201546
'ро'	0.009445443884167399
'ос'	0.009408518771250403
'ли'	0.009219462193115386
'ра'	0.009133795931147957

Топ 15 біграм із пробілами з кроком 1:

symbol	frequency
'_п'	0.018221881514625903
'о_'	0.01776542782216564
'и_'	0.01681750481152214
'_с'	0.01633854107395425
'е_'	0.015106741383342316
'а_'	0.014867884793589138
'_в'	0.014310135898144546
'_н'	0.014137558885652723
'то'	0.011912816094182028
'ст'	0.011480123004890933
'на'	0.011052432147845976
'но'	0.009713084463942034
'_о'	0.009709332789757429
'_к'	0.009544259125634814
'я_'	0.009337917045481547

Топ 15 біграм без пробілів з кроком 2:

symbol	frequency
'то'	0.014414832255310475
'ст'	0.013922964324065635
'на'	0.013094321953409914
'но'	0.011797444945503102
'ен'	0.011110602338809858
'ов'	0.0106807718042986

'по'	0.010441484702611923
'ко'	0.010132774799818615
'ал'	0.010042672866467458
'не'	0.009722146316677277
'ни'	0.009537511207351137
'ро'	0.00944593219312537
'ос'	0.009409005171260141
'ли'	0.009219938819310174
'ра'	0.009134268128582845

Топ 15 біграм із пробілами з кроком 2:

symbol	frequency
'_п'	0.018222929800887196
'о_'	0.01776644984911149
'и_'	0.016818472305423857
'_с'	0.016339481013560583
'е_'	0.015107610458768605
'а_'	0.014868740127839399
'_в'	0.01431095914566963
'_н'	0.014138372204998268
'то'	0.011913501426343521
'ст'	0.011480783444660247
'на'	0.011053067982996434
'но'	0.009713643247786072
'_о'	0.009709891357771477
'_к'	0.009544808197129304
'я_'	0.009338454246326587

Ентропія для символів із пробілами: 4.388710943038407; R=12,24%

Ентропія для символів без пробілів: 4.453045851255678; R=10,94%

Ентропія для біграм з кроком 1 із пробілами: 3.9973064259922904; R=20,06%

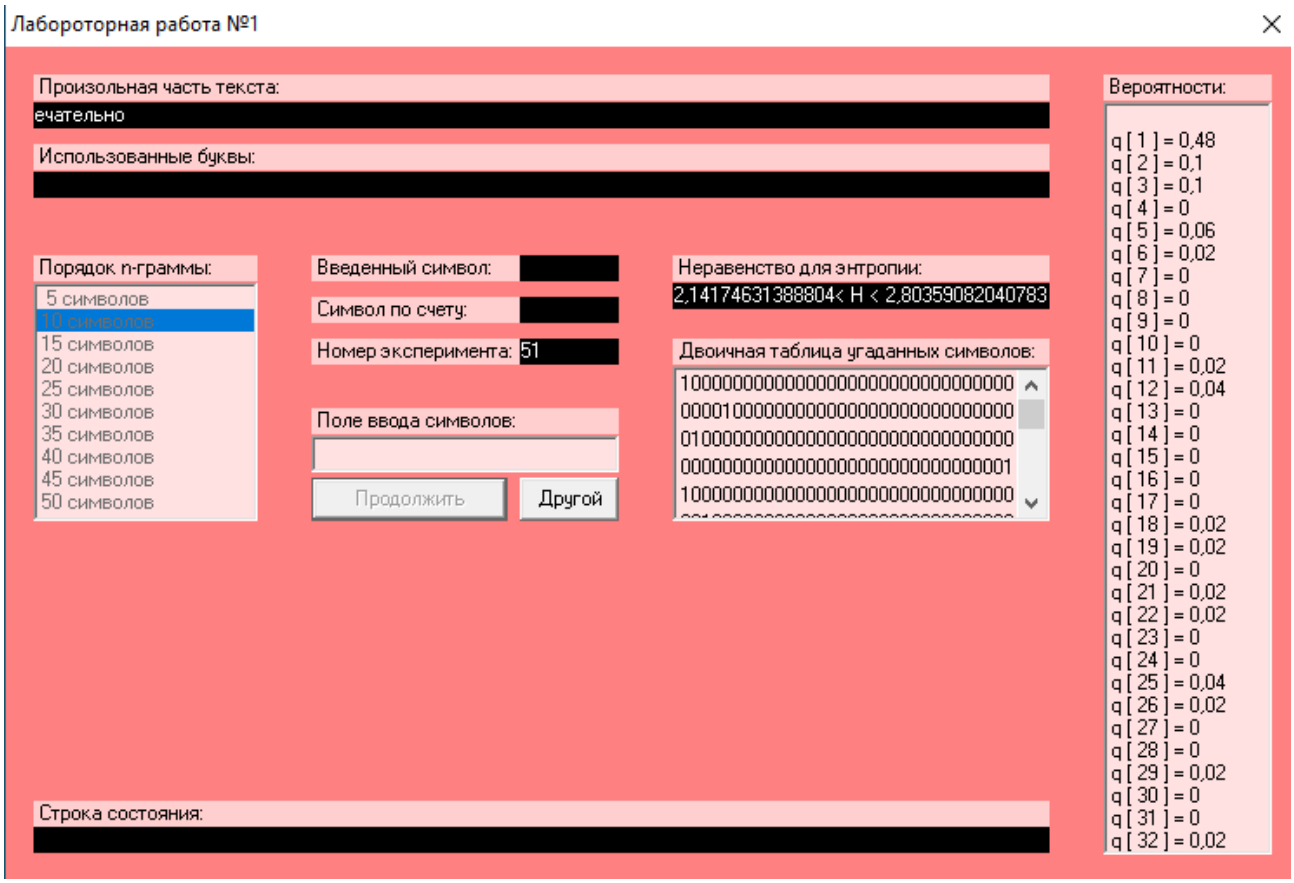
Ентропія для біграм з кроком 1 без пробілів: 4.145286886833119; R=17.1%

Ентропія для біграм з кроком 2 із пробілами: 3.9969700723940513; R=20.08%

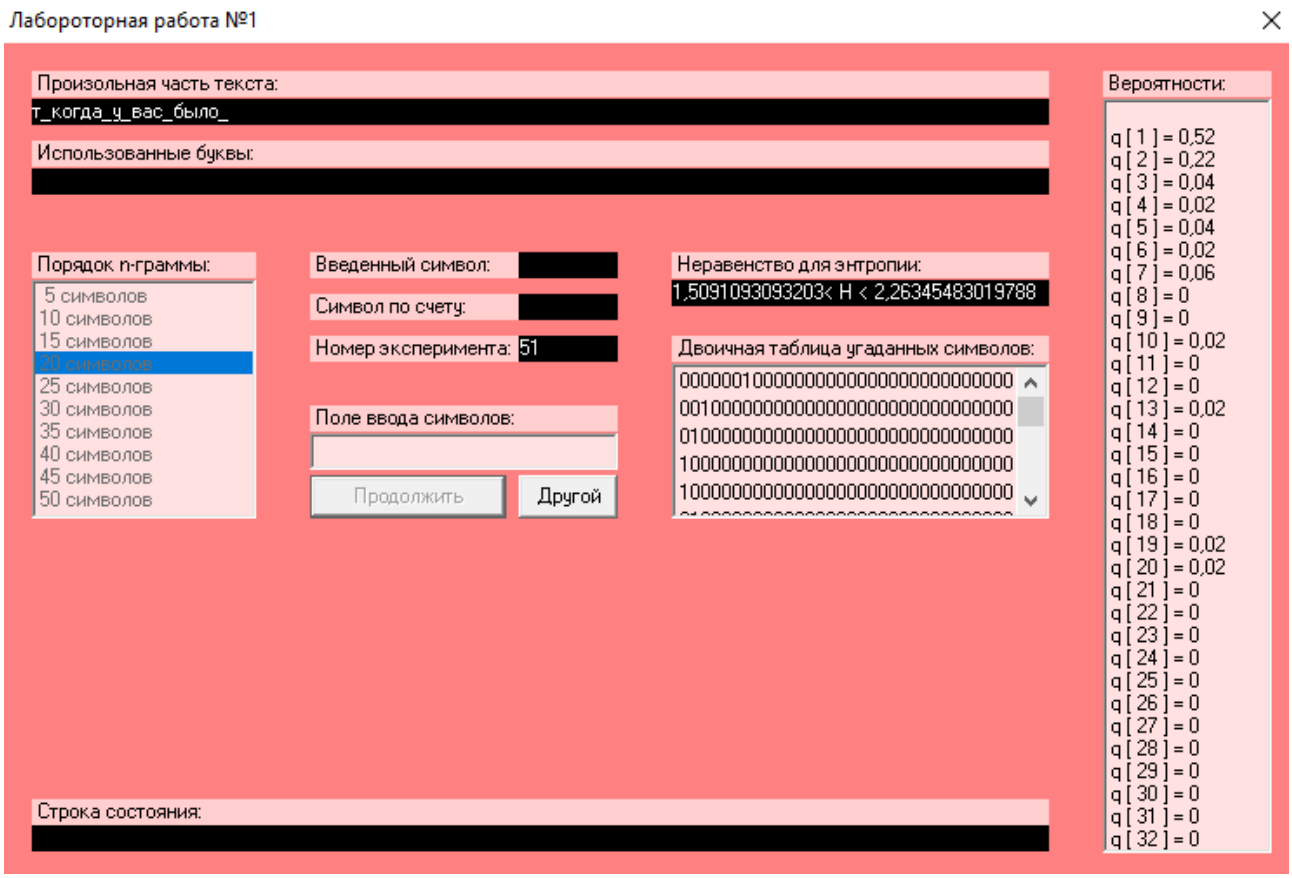
Ентропія для біграм з кроком 2 без пробілів: 4.144983550206557; R=17.12%

2) За допомогою програми CoolPinkProgram оцінити значення $H^{(10)}$, $H^{(20)}$, $H^{(30)}$.

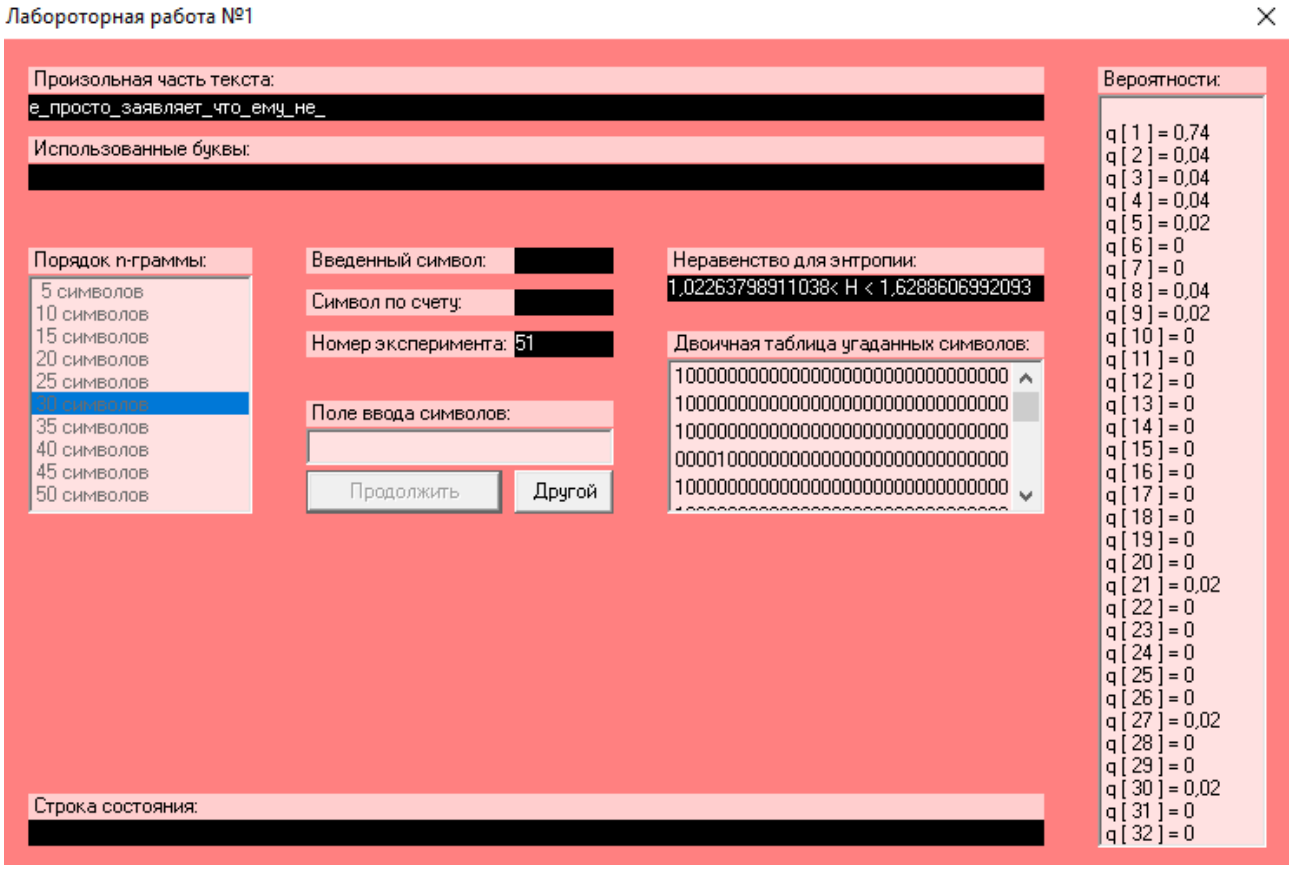
$H^{(10)}$=2,472



$H^{(20)}$=1,886



$H^{(30)}$=1,325



3)Надлишковість:

$R_1 = 1 - \frac{2.472}{5} = 0.5056$

$R_2 = 1 - \frac{1.886}{5} = 0.6228$

$R_3 = 1 - \frac{1.325}{5} = 0.735$

Надлишковість російської мови становить 50,56%; 62,28%; 73,5% відповідно

Висновок: виконавши цю лабароторну роботу ми и ознайомились з поняттями ентропії, частотності і біграм, та отримали перший досвід роботи з ними.Як і очікувалось, було підтверджено, що найчастіше вживані букви російського алфавіту – це “_”, “о” “а” “е”, а найрідше вживані – “ф”, ”щ”, ”э”.Найчастіше вживані біграми – це “_п”, “о_”, “и_” у тексті з пробілами, та “то”, “ст”, “на” у тексті без пробілів.Отримана середня надлишковість російської мови становить 62,11%.